Molecular Phylogenetics and Evolution 67 (2013) 429–435

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Assessing statistical reliability of phylogenetic trees via a speedy double bootstrap method

Aizhen Ren *, Takashi Ishida, Yutaka Akiyama

Graduate School of Information Science and Engineering, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

ARTICLE INFO

Article history:
Received 27 August 2012
Revised 17 January 2013
Accepted 8 February 2013
Available online 26 February 2013

Keywords:
Phylogenetic tree
Maximum likelihood
Tree selection
Bootstrap probability
Double bootstrap
Rapid computation

ABSTRACT

Evaluating the reliability of estimated phylogenetic trees is of critical importance in the field of molecular phylogenetics, and for other endeavors that depend on accurate phylogenetic reconstruction. The bootstrap method is a well-known computational approach to phylogenetic tree assessment, and more generally for assessing the reliability of statistical models. However, it is known to be biased under certain circumstances, calling into question the accuracy of the method. Several advanced bootstrap methods have been developed to achieve higher accuracy, one of which is the double bootstrap approach, but the computational burden of this method has precluded its application to practical problems of phylogenetic tree selection. We address this issue by proposing a simple method called the speedy double bootstrap, which circumvents the second-tier resampling step in the regular double bootstrap approach. We also develop an implementation of the regular double bootstrap for comparison with our speedy method. The speedy double bootstrap suffers no significant loss of accuracy compared with the regular double bootstrap, while performing calculations significantly more rapidly (at minimum around 371 times faster, based on analysis of mammalian mitochondrial amino acid sequences and 12S and 16S rRNA genes). Our method thus enables, for the first time, the practical application of the double bootstrap technique in the context of molecular phylogenetics. The approach can also be used more generally for model selection problems wherever the maximum likelihood criterion is used.

1. Introduction

The analytical methods used in the field of molecular phylogenetics are important basic tools for reconstructing the evolutionary history (phylogenetic relationships) of molecules and organisms. Molecular phylogenetic methods are primarily used in the context of biological systematics, but they find applications in a wide variety of other fields in addition, as diverse as community ecology (Webb et al., 2002), biogeography (Wiley, 1981) and proteomics, including the inference of protein–protein interactions (Pazos and Valencia, 2001) similarity. Many methods of phylogenetic reconstruction have been developed and are in regular use (Felsenstein, 2004). However, those based on maximum likelihood estimation have proved most effective for reconstructing phylogenies using molecular sequence data (DNA, protein, etc.). Early work on this application of maximum likelihood was conducted by Felsenstein (1981), whose approach involved computing the maximum likelihood value for many topologies, and selecting the topology with the highest likelihood (the maximum likelihood (ML) tree) as the most probable candidate for the true topology.

It must be noted that the maximum likelihood values are dependent on the particular characteristics of a random variable: the molecular sequences that constitute the underlying data for phylogeny reconstruction. Thus, some analysis of the statistical reliability of the estimated ML tree or multiple alternative trees should be undertaken. Statistical hypothesis testing is commonly used for this purpose, and the bootstrapping technique is a well-known computational method for calculating reliability when a simple mathematical formula is difficult to derive. Bootstrapping is a resampling method that approximates a random sample by creating a bootstrap sample, generated by random sampling with replacement from the original single data set. In the context of phylogenetic tree selection, Felsenstein (1985) proposed the use of bootstrapping to place confidence intervals on phylogenies. He defined the *p*-value of a tree according to a frequency called the bootstrap probability (BP); the proportion of bootstrap pseudoreplicates of the original data set in which the tree is found to be optimal. However, it is known that under some circumstances the naive bootstrap probability can be biased (e.g., Hillis and Bull, 1993; Sanderson and Wojciechowski, 2000). Thus, some advanced bootstrap methods have been proposed, to achieve higher accuracy (Hall, 1992; Efron et al., 1996; Efron and Tibshirani, 1998; Shimodaira, 2002). Among these, the double bootstrap (Hall, 1992; Efron and Tibshirani, 1998) has been shown to be third order accurate

* Corresponding author. Fax: +81 3 5734 3646.
E-mail address: ren@bi.cs.titech.ac.jp (A. Ren).

1055-7903 © 2013 Elsevier Inc. Open access under CC BY-NC-ND license.
http://dx.doi.org/10.1016/j.ympev.2013.02.011

and may hold great potential as a measure of phylogenetic tree support. However, the method imposes huge computation burdens and has yet to be applied in the context of molecular phylogenetics. To overcome this computational difficulty we propose a speedy double bootstrap method to compute the reliability of phylogenetic trees. For comparison, we also developed a procedure to implement the regular double bootstrap (Hall, 1992; Efron and Tibshirani, 1998), and we used these methods to analyze the mammalian mitochondrial protein sequences and genes for 12S and 16S rRNA. To illustrate the utility of our speedy double bootstrap, we compared results from this method with those from the regular double bootstrap, the traditional bootstrap proportion (BP), and the multiscale bootstrap technique (AU test) described by Shimodaira (2002).

2. Materials and methods

2.1. The double bootstrap method

In this study, homologous sites of aligned molecular sequence data are regarded as the units of sampling, and we use DNA data as the example for the following methodological descriptions. Suppose we have m homologous sequences, each with n nucleotide sites. These data can be represented as a $m \times n$ matrix $\mathbf{X} = \{x_{jh}\} = \{\mathbf{x_1},\ldots,\mathbf{x_n}\}$, where $\mathbf{x}_h$ is the value of the h-th site and x_{jh} is one of the four deoxyribonucleotides (T, C, A, or G).

$$\begin{aligned} &Species1 : x_{11}\ x_{12}\ \cdots\ x_{1n} \qquad (1)\\ &Species2 : x_{21}\ x_{22}\ \cdots\ x_{2n}\\ &\ \vdots\ \ \vdots\\ &Speciesm : x_{m1}\ x_{m2}\ \cdots\ x_{mn} \end{aligned}$$

The log-likelihood can be expressed as

$$l(\theta;\mathbf{X}) = \sum_{h=1}^{n} logf(\mathbf{x}_h;\theta) \qquad (2)$$

where $f(\mathbf{x}_h;\ \theta) = f(x_{1h}, x_{2h},\ldots, x_{mh};\ \theta)$ is the probability that at a particular homologous site, species 1 has base x_{1h}, species 2 has x_{2h} and species m has x_{mh}. The vector θ denotes unknown parameters such as the edge lengths (branch lengths) of a tree, and the base substitution rates along these branches. Here we assume that the base substitution rates have already been estimated, so θ denotes only the unknown edge lengths. For a given tree topology, θ is estimated by maximizing the log-likelihood, and the maximum log-likelihood of any tree topology i is given by

$$l_i(\hat{\theta}_i;\mathbf{X}) = \sum_{h=1}^{n} logf_i(\mathbf{x}_h;\hat{\theta}_i) \qquad (3)$$

The topology with the highest value of $l(\hat{\theta};\ \mathbf{X})$ is the maximum likelihood phylogeny (T_{ML}) for data set $\mathbf{X}$, and is thus the most likely candidate for the true topology. To define null hypotheses for performing model comparisons, we must first recognize that molecular sequence data are discretely distributed; the true distribution for a random variable $\mathbf{x}$ can be expressed as

$$q(\cdot) = \{q(\mathbf{x}_1), q(\mathbf{x}_2),\ldots, q(\mathbf{x}_s)\} \qquad (4)$$

where $s = 4^m$and is the expectation of $l_i(\hat{\theta}_i;\ \mathbf{X})$ with respect to $q(\cdot)$, $i = 1, \ldots, K$, i.e.

$$\mu_i = E_q[l_i(\hat{\theta}_i;\mathbf{X})] = \sum_{h=1}^{n} E_q[logf_i(\mathbf{x}_h;\hat{\theta}_i)] = nE_q[logf_i(\mathbf{x};\hat{\theta}_i)] \qquad (5)$$

where $E_q[logf_i(\mathbf{x};\hat{\theta}_i)] = \sum_{t=1}^{s} q(\mathbf{x_t}) logf_i(\mathbf{x}_t;\hat{\theta}_i)$, for each $i = 1, \ldots, K$. So if we assume that tree T_1 is the best topology, the null and alternative hypotheses will be

$$H_1 : \mu_1 = max_{i=1,\ldots,K}\mu_i \ \ vs.\ H_1^A :\ \ others \qquad (6)$$

and we must continue performing these comparisons as many times as is necessary, assuming in turn that tree T_i, $i = 2, \ldots, K$ is the best topology. Note that the null hypothesis H_1 involves multiple comparisons with the "best" topology (Hsu, 1981): as can be seen from (6), the null contains $k-1$ hypotheses such that

$$H_{1j} : \mu_1 \geqslant \mu_j,\ j = 2,\ldots,K, \qquad (7)$$

The null hypothesis H_1 is a polyhedral convex cone and $\partial\ H_1$, which is boundary of H_1 is nonsmooth at the vertex as well as on the faces of dimensions less than $K-1$. Shimodaira and Hasegawa (1999) proposed a multiple comparisons procedure (the SH-test) to test H_1, but this was shown to be overly conservative and a different method was designed (the AU test), which uses a multiscale bootstrap technique to obtain third-order accurate p-values for testing the null hypothesis. Other authors (e.g., Hall, 1992; Efron and Tibshirani, 1998) had previously developed a double bootstrap method that was also able to provide third-order accurate p-values, but due to high computational requirements this method has not been adopted for phylogenetic applications.

At this juncture it is necessary to briefly review the double bootstrap method. The third-order accurate p-values was first proposed by Efron (1985) for the multivariate normal model, which can be represented as

$$\mathbf{Y} \overset{i.i.d.}{\sim} N_t(\eta, I_t) \qquad (8)$$

This normal model is a simplification of reality. Let $\mathcal{H} \subset \mathbb{R}^t$ be an arbitrarily-shaped region with smooth boundaries denoted by $\partial\mathcal{H}$. We want to calculate a p-value $p(y)$ for testing the null hypothesis $\eta \in \mathcal{H}$. According to Efron (1985), when the true parameter η is on the boundary surface $\partial\mathcal{H}$, the third-order accurate p-value can be expressed as

$$p(y) = 1 - \Phi(d-c) \qquad (9)$$

where d is the signed distance from y to $\hat{\eta}(y)$, with a positive or negative sign when y is, respectively, outside or inside $\mathcal{H}$. The point $\hat{\eta}(y)$ is the closest point to y (in Euclidean distance) on the surface $\partial\mathcal{H}$, and c in formula (9) is a quantity related to the curvature of $\partial\mathcal{H}$ at point $\hat{\eta}(y)$. The double bootstrap method of Hall (1992) and Efron and Tibshirani (1998) begins with a first tier of bootstrap resampling from the multivariate normal model with distribution

$$\mathbf{Y}^* \overset{i.i.d.}{\sim} N_t(\hat{\eta}(y), I_t) \qquad (10)$$

A second tier of resampling is carried out for each of these vectors $\mathbf{Y}^*$, as well as for $\mathbf{Y}$, with the following distributions

$$\begin{aligned} &\mathbf{Y}^{**} \overset{i.i.d.}{\sim} N_t(\mathbf{Y}^*, I_t) \qquad (11)\\ &\mathbf{Y}^{**} \overset{i.i.d.}{\sim} N_t(\mathbf{Y}, I_t) \end{aligned}$$

The second tier quantities in each case are as follows

$$\tilde{p}^* = P(y^{**} \in \mathcal{H}; y^*);\ \tilde{p} = P(y^{**} \in \mathcal{H}; y) \qquad (12)$$

Then, according to Hall (1992) and Efron and Tibshirani (1998), the third-order accurate p-value (9) obtained by the double bootstrap method can be expressed as

$$1 - \Phi(d-c) = P(\tilde{p}^* < \tilde{p};\ \hat{\eta}(y)) + O(n^{-3/2}) \qquad (13)$$

Although the double bootstrap has third-order accuracy, formula (13) suggests that it requires enormous numbers of bootstrap pseudoreplicates (many more than would be practically feasible in most cases), and in addition, computation of $\hat{\eta}(y)$ is known to be difficult. However, we propose a manipulation of the regular double bootstrap that will greatly speed its implementation and thus facilitate its application to real phylogenetic problems. Our method relies on use of formula (14) below (Efron and Tibshirani, 1996),